63-3-2

STOCHASTIC MODELS FOR THE INTERPRETATION OF METEOROLOGICAL DATA

By

F. N. David

University of California

Berkeley 4, California

Final Report

Contract No. AF 19(604)-8051

Project 8624

Task 862402

Date of Report:

January, 1963

Prepared for

GEOPHYSICS RESEARCH DIRECTORATE

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES

OFFICE OF AEROSPACE RESEARCH

UNITED STATES AIR FORCE

BEDFORD, MASSACHUSETTS

Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

Department of Defense contractors must be established for ASTIA services or have their "need-to-know" certified by the cognizant military agency of their project or contract.

All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C.

STOCHASTIC MODELS FOR THE INTERPRETATION OF METEOROLOGICAL DATA

By

F. N. David

1. The Statement of Work to be accomplished under the contract was set out as follows:

Item 1. Discover and describe one or more chance mechanisms or stochastic processes for the behavior of atmospheric variables, and similar random variables, which would lead to the observed serial and spatial correlations of winds, temperatures, pressures, precipitation, etc.

Item 2. Develop optimum procedures for estimating the parameters of the mechanisms, and for determining their significance.

2. Work began on the contract on June 1, 1961. At the behest of Dr. Arnold Court I visited, with an introduction from him, Mr. C. S. Durst, then recently retired from the Meteorological Office. Mr. Durst was widely known for his work on winds and had published a great many important papers on the topic.* He was generous of his time and in addition to making known to me all the available literature he told me in detail, during the first and subsequent visits, such ideas as he had for further research on these problems. That I did not take advantage of his generosity was solely because my own research developed along different lines.

3. In mid-June, having completed the gathering of information in London as far as time allowed--I was not able to take advantage of an invitation from the Director to visit the Royal Meteorological

---

*It had been my hope to collaborate with him on my return to England but he died in December, 1961.

Office--I visited with Dr. Arnold Court then chief of the Applied Climatology Branch in Waltham, Massachusetts. We spent one whole working day together during which he reinterpreted the Statement of Work and asked for an empirical interpretation of spatial correlations between pressure measurements. This, then, was the first problem which I tackled. [Report AFCRL-62-461(I).]

4. Pressure as a Function of Distance I. [AFCRL-62-461(I)] The raw material of the empirical investigation consisted of the daily pressures at stations in the United States and Canada for the three months January, February, and March of the years 1949-1953. The heights at which the pressures were 500 m.b. were found and the correlation coefficient between these heights, taking each station in turn with every other station, were calculated. Enough stations were taken to enable contour lines to be drawn. An empirical surface to describe the space-correlation surface thus formed was requested.

Using parallels of latitude as delineating possible cross sections of the surface the functional form of such cross sections was investigated. The various functional forms which were tried and the variety of methods of fitting them are described in the report, with the manipulative mathematics necessary. It is enough here to note that one was driven to the inescapable conclusion that the functional form must consist of a sum of damped harmonics and that the width of the United States was not sufficient to allow estimation of the parameters involved. Thus if a station X is chosen as origin, and the correlation of 500 m.b. heights $r_{x,i}$ with a station Y_i $(i=1,2,\cdots)$ on the same latitude and a distance

d_i from X is calculated, then an estimate of this correlation may be obtained from

$$(1) \qquad \tilde{r}_{x,i} = \sum_{j=1}^{p} e^{-a_j \varphi(d_i)} A_j \cos(\beta_j d_i + \gamma_j)$$

where a_j, A_j, β_j, γ_j are constants to be determined and $\varphi(d_i)$ is equal to d_i or d_i^2. It is clearly not enough to take $p = 1$ but how large p should be it is not possible to say, neither is it possible to determine what φ should be.

Given that one may eliminate by trial and error the common statistical functions used to describe correlation data it was found that no estimates of the damped harmonics parameters are possible since the distance between the two stations farthest apart was not large enough to cover a complete period. The general pattern of the beginning of the space correlation data was very like the pattern of the serial time correlations investigated by Gilbert Walker--like enough to enable one to be reasonably sure that (1) was appropriate--but his correlations oscillated about -0.2 whereas it was not possible to see what the space correlations did.

Early in the investigation it became clear that the empirical approach involving choosing an arbitrary functional form to graduate the space-correlogram was inappropriate for the problem. It was a good idea but the data are such that it did not work out. Later in the year a new approach was adopted.

5. Pressure as a Function of Distance: II. [AFCRL-62-1012] The problem of relating the correlation of pressure measurements at two stations with the distance between them was investigated from

one aspect in the first paper. It is suggested that a more fruitful approach to the interpretation of such correlations is by building a stochastic model. The model delineated in Pressure as a Function of Distance: II is almost certainly too simple to describe what is, after all, a very complex state of affairs, but time did not allow the appropriate generalizations to be made. Such results as are obtained, however, suggest that this method of attack on the problem is entirely suitable for the purposes of generating a correlation surface such as was described as being required in Pressure as a Function of Distance: I. Generalizations both of model and of the method of approach to the problem are suggested in a further section of this report.

It is certain that the position of the estimated high pressure center nearest the geographical center of the United States can be marked on a map at a fixed time each morning (say). If we wish to study, as Court did, pressures for the months of January, February and March, we would wish to know the variation of the position of the high pressure center for these three months for a number of years. We may calculate the average of all these positions and use this as a center of coordinates with axes of reference the parallel of latitude and longitude running through this center. (It is plausible to suggest that this average position will be not very different from the geographical center, but if it is found not to be so, but varies (say) according to the season of the year, this could be allowed for.) It will be assumed for the purposes of this present model that the daily high pressure center is distributed about the average center as in a normal bivariate surface, the

variability along the line of latitude being σ_1, along the line of longitude being σ_2, and with correlation ρ. For any day σ_1, σ_2 and ρ are constants but in the over-all picture they are regarded as random variables each having a p.d.f.

Consider any station of known geographical position, and suppose that the pressure recorded at the fixed time on a given day is built up from a basic pressure, here regarded as constant throughout, plus a constant multiple, c, of a random quantity. This quantity is supposed to depend on the distance of the station from the high pressure center recorded for the fixed time on the given day. Thus if (X, Y) are coordinates of the high pressure center referred to the average position, and (U, V) are the coordinates of the station referred to the same origin and axes, p is the actual pressure and p_0 the basic pressure, we assume that

$$p - p_0 = \delta_p = \frac{c}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[(U-X)^2 + (V-Y)^2\right]\right\}.$$

σ will be treated as variable. For fixed σ_1, σ_2, ρ and σ, δ_p can be averaged over all possible values of (X, Y) and then assuming that σ_1^2, σ_2^2 and σ^2 are distributed as gamma variables with

$$\mathcal{E}(\sigma_1^2) = b, \quad \mathcal{E}(\sigma_2^2) = c, \quad \mathcal{E}(\sigma^2) = a$$

and that

$$p(\rho) = \tfrac{3}{4}(1-\rho)(1+\rho), \qquad -1 < \rho < +1,$$

the average of δ_p is obtained allowing for these variations. The necessary integrations were difficult and the answer was obtained

as a series expansion, attention being paid to relative orders of magnitude. Var(p) may similarly be deduced and also the covariance between the pressure measurements recorded at two different stations.

It was supposed that $a = \alpha b = \beta c$ where α and β are small and less than unity. Further, U and V were standardized and we write

$$U^* = \frac{U}{\sqrt{b}}, \qquad V^* = \frac{V}{\sqrt{c}}$$

so that

$$R^2 = U^2 + V^2, \qquad R^{*2} = U^{*2} + V^{*2}.$$

R will thus be the actual distance of the station from the average position of the high pressure centers and R^* the standardized distance. If we consider two stations with coordinates (U_1, V_1) and (U_2, V_2) it was shown, to a first order of approximation, that we have, for the correlation between the pressure measurements at two stations,

$$\frac{1}{W_1 W_2}\left[\exp\left\{-\frac{1}{2}\left[\frac{(U_1^* - U_2^*)^2}{\alpha(\alpha+2)} + \frac{(V_1^* - V_2^*)^2}{\beta(\beta+2)}\right]\right\}\right.$$

$$\left. - \frac{\sqrt{\alpha\beta}\sqrt{(2+\alpha)(2+\beta)}}{(1+\alpha)(1+\beta)} \exp\left\{-\frac{1}{2}\left[\frac{U_1^{*2} + U_2^{*2}}{(1+\alpha)(2+\alpha)} + \frac{V_1^{*2} + V_2^{*2}}{(1+\beta)(2+\beta)}\right]\right\}\right]$$

where

$$W_1^2 = 1 - \frac{\sqrt{\alpha\beta}\sqrt{(2+\alpha)(2+\beta)}}{(1+\alpha)(1+\beta)} \exp\left\{-\left[\frac{U_1^{*2}}{(1+\alpha)(2+\alpha)} + \frac{V_1^{*2}}{(1+\beta)(2+\beta)}\right]\right\}$$

and similarly for W_2. So far b and c and therefore α and β have been supposed different. Some simplification results if we

write $\alpha = \beta = \delta$ so that, say,

$$R_1^{*2} = U_1^{*2} + V_1^{*2} = (U_1^2 + V_1^2)/b = D_1^2; \quad R_2^{*2} = D_2^2$$

$$(U_1^* - U_2^*)^2 + (V_1^* - V_2^*)^2 = D^2.$$

Thus D_1 and D_2 are proportional to the distances of the two stations from the origin and D to the distance between them. We have then as our first approximation,

$$r_{12} = \frac{\exp\left\{-\frac{D^2}{2\delta(2+\delta)}\right\} - \frac{\delta(2+\delta)}{(1+\delta)^2}\exp\left\{-\frac{(D_1^2 + D_2^2)}{2(1+\delta)(2+\delta)}\right\}}{\prod_{i=1}^{2}\left[1 - \frac{\delta(2+\delta)}{(1+\delta)^2}\exp\left\{-\frac{D_1^2}{(1+\delta)(2+\delta)}\right\}\right]^{1/2}}.$$

6. Application of Deduced Formula Correlation. So far the model has been built and the consequences deduced without reference to the data of experience. Time did not permit a data check of the foundations of the model. We proceeded therefore to guess values for b and for δ and for the origin of coordinates. We took therefore, entirely arbitrarily, the origin of coordinates (i.e., the average position of the high pressure centers) at 40°W, 100°N, and for the sake of example calculated the distances of stations 562, 553, 451, 445, 764 and 662 from this center, the distances being those on the sphere. The distances of all the stations from 562 were also found. Again guessing we chose $\alpha = \beta = \delta = 0.1$ and $\sqrt{b} = 18°$. From our formula for r_{12} we then worked out the correlations to be expected between the pressure measurements at station 562 and all the other stations. These correlations should not be directly comparable with those calculated by Dr. Cooley from actual

data in that his correlations were, I believe, based on average daily measurements, but they should be reasonably in accord with his. The correlations based on our formula and the ones actually observed are shown in the table below:

Table: Correlations of Pressure Measurements With Those of Station 562

Station	553	451	445	764	662
Actual correlation	0.81	0.89	0.53	0.74	0.86
Formula correlation	0.81	0.90	0.50	0.75	0.87

7. Discussion of Results. The stations chosen were those near 562 because it was anticipated that given the approximations involved, and there were many, the correlations would be too large when the distances between the stations became large. Again when the distances from the average position became large it is to be expected that the correlations will be overemphasized. Further calculations using different groups of stations showed this to be indeed the case. Thus while the results obtained are somewhat remarkable in that those produced by the model without reference to real data agree closely with those actually obtained from real data this can, I think, be no more than an indication of what might be done by building further models of this kind. The first thing which clearly it is important to do, is to obtain some idea of the average position of the high pressure centers, and of the other parameters involved in the model. Only then is it possible to say

whether the model is really serviceable and worthy of further development.

8. Further Possible Research. Emphasis has been laid on the fact that the stochastic model proposed above is too simple to describe the pressure complex, and moreover that the approximations made in working out the algebra of the model are such as to render it only of rudimentary value. It is noted immediately above that for stations close to what is anticipated to be the average center of the high pressure system the model is useful, but what is required is a model which will be adequate for prediction over a very wide range. Clearly the numerical investigations noted as necessary in section 7 should be the first part of any further research. Given that some idea is thus obtained then the following are a few of the ways in which further mathematical attack appears possible.

(i) A modification of the model given here can be made by introducing more variability and assuming that the average position of the high pressure centers varies about (say) the geographical center of the United States.

(ii) The constant c may be assumed to be different for different stations.

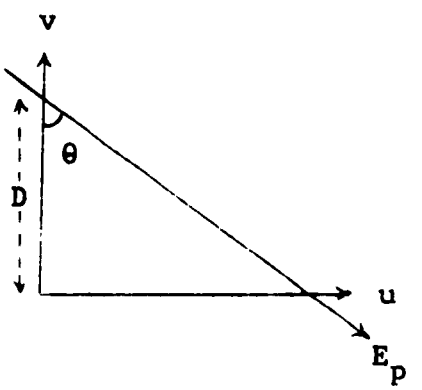
(iii) The isobaric contours suggest not a normal bivariate surface, as was assumed with the present model, but rather a double Edgeworth surface. (Probably this is the reason why the present model is inadequate for large distances.) The mathematical analysis involved in introducing a double Edgeworth series in this way is not difficult although it is laborious. As many parameters as desired can be introduced into the model by taking successive terms of the

series. For the type of approach to the problem of pressure as a function of distance as illustrated in AFCRL-62-1012 the assumption of the double Edgeworth series will probably give as general a result for the correlations as can be expected.

(iv) A different method of approach to the pressure measurement problem appears possible along the lines of that used for the "effective" stochastic model for precipitation (see below). One could assume the center of the high pressure system traveling along a path at a given rate, the various descriptive quantities intervening being assumed to be random variables with given p.d.f.'s. In this way a stochastic model could be built up and if required the further refinement of there being several high pressure centers could be added. I had a preliminary look at such a method of approach and it appears entirely possible.

9. An "Effective" Stochastic Model for Interpreting the Correlation Between the Precipitation at Two Stations. [AFCRL-62-495] In this paper I was concerned with interpreting precipitation data as actually measured on the ground. To this end I introduced a concept, which although not known to me before must have been invented many times previously, that of the "effective" path of a rainstorm. Suppose any number of stations, spread over a large area, all measuring the precipitation from a rainstorm simultaneously at certain fixed points of time. At any given time point, knowing the geographical positions of the stations and the amount of precipitation at each we may calculate the geographical position of the center of gravity of the precipitation from the several stations, and we may do this for each time point. If the rainstorm is moving

then the center of gravity of measured precipitation will be different for each time point, and the best fitting curve to these centers of gravity will give what I have called the effective path of the rainstorm and will enable the rate at which the storm is moving to be estimated. I have assumed the effective path to be linear but no intrinsic difficulty is introduced if any other functional form is envisaged. It would be desirable for numerical work to be carried out to investigate the actual effective paths of rainstorms and their relationship, if any, with the low pressure center of the storm.



If linearity is assumed then the effective path E_p of the storm can be described by two quantities. Consider any axes of reference (u, v) and any origin of coordinates. Let E_p cross the v axis at a distance D from the origin of coordinates at an angle θ as shown. These descriptive quantities θ, D will be different for different storms and it will be supposed that for all storms

$$p(D) = \frac{1}{D_1 + D_2}\, dD, \qquad -D_1 < D < D_2,$$

$$p(\theta) = \sin\theta\, d\theta, \qquad 0 < \theta < \pi.$$

The center of gravity of the storm which we will speak of as the effective center may for any particular storm be supposed to be moving at a constant rate of r miles per unit of time, and for all storms it will be assumed that

$$p(r) = \frac{1}{\Gamma(f+1)} r^f e^{-r} dr, \qquad 0 < r < +\infty.$$

Now let us consider the composition of the storm. We build this up from the effect on the ground where the precipitation will be assumed finally to arrive as drops of water. Suppose each storm is composed of a number of showers. Each of these showers will be assumed to have an effective center, i.e., a center of gravity of the drops which compose it. These centers of gravity (or effective centers) may be supposed to be distributed in bivariate normal fashion, with center the effective storm center, the minor axis of the elliptic contours of equal density being the effective path of the storm, and the standard deviations in the directions of the major and minor axes being Σ_2 and Σ_1. Further, given any particular effective <u>shower</u> center the drops composing this shower are assumed to have a spatial bivariate normal distribution with center the shower center, with standard deviations along the major and minor axes of equal density ellipses equal to σ_2 and σ_1 and with minor axis parallel to the effective path of the storm. When drops fall on the ground it is supposed that all the drops composing one shower have fallen simultaneously out of the storm and that they fall independently of one another. Further, any shower will be assumed to fall independently of any other shower. We have thus set up a "random" mechanism to describe the observed ground effects.

It remains to characterize the intensity of the storm. We will suppose that all rain drops are of the same size for convenience. (There is no intrinsic difficulty in introducing a drop size distribution into the model.) If one particular shower, say the ith, has d_i drops, we will let

$$p(d_i) = \frac{\eta^{d_i} e^{-\eta}}{d_i!}, \qquad d_i = 0,1,2,\cdots,\infty .$$

η will thus represent the average number of drops in a shower for a given storm. If the probability that a shower falls out of the given storm in time dt is λdt, and not more than one shower is allowed to fall in this given instant of time, λ will represent the average number of showers which fall from the given storm in the fixed time. λ and η will therefore together represent the intensity of the given storm. We shall assume λ and η are independent of each other and of r the rate at which the effective storm center is moving. (An obvious generalization of the model is to assume that they are all correlated in some way.) λ and η will be different for different storms and in the absence of further practical information we will assume

$$p(\eta) = \frac{6}{A^3}\,\eta(A-\eta)\,d\eta, \qquad 0 < \eta < A,$$

$$p(\lambda) = \frac{6}{B^3}\,\lambda(B-\lambda)\,d\lambda, \qquad 0 < \lambda < B,$$

where A and B are two positive numbers with, it may reasonably be supposed if required, $A > B$.

It is desired to compute the correlation between the rainfall

at two stations. This can be done for a particular storm or, in line with the data the writer was given for pressure measurements, by forming a bivariate table of the amounts of precipitation at two stations for a given period of a year over a number of years, and from this table computing the correlation. The model given can be used for either case. Let the two stations be W_1 and W_2 and let the catchment (or target) areas in each be of area H_1^2 and H_2^2, respectively. If the geographical coordinates of W_1 and W_2 are (U_1, V_1) and (U_2, V_2), referred to the given axes of coordinates, with

$$U_2-U_1 = R_1, \quad V_2-V_1 = R_2, \quad R^2 = R_1^2+R_2^2, \quad R_z = R_1 \sin\theta - R_2 \cos\theta$$

then R is the distance between the two stations. The required correlation is

$$\rho_{12} = \frac{\frac{3}{40\cdot\sigma_1\sigma_2\cdot B}\frac{\mathcal{I}}{\pi} - \frac{\pi^2}{8f(D_1+D_2)}}{\left[\left\{\frac{3}{40\cdot\sigma_1\sigma_2 B} - \frac{\pi^2}{8f(D_1+D_2)} + \frac{\pi}{2H_1^2AB}\right\}\left\{\frac{3}{40\cdot\sigma_1\sigma_2\cdot B} - \frac{\pi^2}{8f(D_1+D_2)} + \frac{\pi}{2H_2^2AB}\right\}\right]^{1/2}}$$

where

$$\mathcal{I} = 2\int_0^1 \frac{1}{\sqrt{1-z^2}} \exp\left\{-\frac{1}{4}\left[\frac{R^2z^2}{\sigma_1^2} + \frac{R^2(1-z^2)}{\sigma_2^2}\right]\right\} dz.$$

When R is large, $\mathcal{I}$ will not be very different from zero, and the correlation between the precipitation at the two stations will be negative. When R is small there will be a positive correlation. An explicit expression for $\mathcal{I}$ does not appear possible except in special cases but two expansions can be evolved which will be adequate.

a) R <u>small and</u> $< 2\sigma_1$.

It was assumed that σ_1 was smaller than σ_2 and this expression will be useful only if $R < 2\sigma_1$, which will imply that the stations are fairly close together. We have

$$\mathcal{I} = 2\int_0^1 \frac{1}{\sqrt{1-z^2}} \exp\left\{-\frac{1}{4}\left[\frac{R^2z^2}{\sigma_1^2} + \frac{R^2(1-z^2)}{\sigma_2^2}\right]\right\} dz.$$

Let

$$\frac{1}{s^2} = \frac{R^2}{4}\left[\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right]$$

so that

$$\mathcal{I} = 2\left[\exp\left\{-\frac{1}{4}\frac{R^2}{\sigma_2^2}\right\}\right]\int_0^1 \frac{1}{\sqrt{1-z^2}} \exp\left\{-\frac{z^2}{s^2}\right\} dz$$

or

$$\mathcal{I} \doteq \pi\left[\exp\left\{-\frac{R^2}{4\sigma_2^2}\right\}\exp\left\{-\frac{1}{2s^2}\right\}\right]\left\{1 + \frac{1}{16s^4} + \frac{1}{4}\left(\frac{1}{16s^4}\right)^2\right\}.$$

b) R <u>large</u>.

Write

$$G = 1 - \exp\left\{-\frac{1}{s^2}\right\}.$$

Then

$$\mathcal{I} = 2\left[\exp\left\{-\frac{R^2}{4\sigma_2^2}\right\}\right]\left[\sin^{-1}(s\cdot G^{1/2}) + s\cdot G^{3/2}\left\{-\frac{1}{12} + \frac{G}{5}\left(\frac{s^2}{8} - \frac{7}{96}\right)\right.\right.$$

$$+ \frac{G^2}{7}\left(\frac{9}{32}s^4 + \frac{13}{192}s^2 - \frac{5}{128}\right)$$

$$\left.\left. + \frac{G^3}{9}\left(\frac{25}{64}s^6 + \frac{171}{168}s^4 + \frac{35}{768}s^2 - \frac{787}{30720}\right) + \cdots\right\}\right].$$

These two expressions will be sufficient to cover the range of possible values. Given the values of the several constants entering into the formulae the value of the correlation between the precipitations at two stations can accordingly be predicted. The formulae were rewritten to take account of situations where one target area overlapped the other or alternatively one was contained within the other. It will be noted that if one station is taken as center it is possible, by varying R to draw geographical contours on which the correlation is constant, in other words, to draw a correlation surface for precipitation data such as was envisaged for pressure data.

10. Practical Considerations Concerning the "Effective" Stochastic Model for Precipitation. Access was not available to sufficient data to enable a complete trial of the propriety of the model suggested to be made. The first and most important point is to investigate what was called the effective path of the storm. I was not able to do this. I did however have made available to me for California only, some paths of low pressure centers which were calculated for the Santa Barbara weather project. These data were rather scanty but from them it was possible to decide that assumptions of rectangularity for the p.d.f. of D, of $\sin\theta$ for the p.d.f. of θ, and of a gamma distribution for r, were not unreasonable. It would seem that numerical values for σ_1 and σ_2 should not be difficult to obtain provided it is possible to measure the precipitation over a short enough period of time, say 5 or 10 minutes, for a group of stations. σ_1 and σ_2 will undoubtedly be variable in practice, but an average value of these will probably

be good enough. If they are found to be of great variability then the model will need to be extended to take account of this. There remains to find estimates of B and of AB. It is possible to obtain these either from the raw data, or which is possibly less complicated, to take two pairs of stations the distances between the elements of each pair being small but the distance between the pair being moderately large. If the actual correlations between observed data are now calculated, and the formula given here for the correlation is used, one may back solve to get values for B and for AB and consequently for A.

11. Further Possible Work Following from the "Effective" Model. The acid test of any model is how well it accords with the data of experience and until this actual numerical data is analyzed with the objective of testing the several assumptions which are made, there is little point in further generalization of the model. Modifications to accord with numerical results will not be difficult to make once it is known what is required. One very obvious generalization which may be necessary--as stated immediately above--is to allow for variability in the scale parameters of the "drop" distribution. This may be done fairly simply by assuming for each of the p.d.f.'s of σ_1 and σ_2 an inverse gamma (Pearson Type V) distribution but whether the resulting integrations can then be performed is something which has yet to be investigated.

12. Persistence in a Chain of Multiple Events when there is Simple Dependence. [AFCRL-62-496] During one of my visits to Mr. Durst we discussed the problem of persistence in weather and he drew my attention to several papers in the Proceedings of the

Royal Society in which this problem was treated. I studied these papers but they seemed old fashioned from a statistical point of view and I decided to see what could be done using a more modern method of attack. The problem is in essence a relatively simple one. Given a series of observations which are consecutive in operational time, it is required to describe the correlation between any pair of them. As research developed on this problem it was found more rewarding to work with a persistence factor, θ, which actually attempted to describe the dependence of each observation on the one which precedes it, and which is connected with the correlation in a very simple way.

Suppose a sequence of n consecutive intervals of operational time; during each interval one only of s mutually exclusive events E_i $(i=1,2,\cdots,s)$ must happen. Thus for example we might describe rainfal in the terms heavy (E_1), medium (E_2), light (E_3), trace (E_4) and none (E_5). These would be five mutually exclusive possibilities. The operational time will be interpreted according to what we are studying. It might be consecutive actual time intervals (hours, minutes, days) at the same weather station or it could be distances with n weather stations instead or a combination of the two or anything else in which we are interested. The chain of events is assumed to have reached the equilibrium state so that

$$P(E_i) = p_i, \quad i=1,2,\cdots,s \quad \text{and} \quad \sum_{i=1}^{s} p_i = 1.$$

If there is no dependence between events, i.e., if for example the direction from which the wind blew at 10 a.m. was uncorrelated

with the direction from which the wind blew at 9 a.m., then

$$P\{E_i|E_i\} = p_i = P\{E_i|E_j\}, \qquad i=1,2,\cdots,s;\ j=1,2,\cdots,s.$$

We imagine that there will be a correlation of some kind, however, and allow for it in the following way. Since n consecutive time intervals are supposed, we have a chain of n consecutive events. At any point in the chain if we take (say) the kth and $k+1$-st positions, we write

$$P\{E_i \text{ in the } (k+1)\text{st position given } E_i \text{ in the } k\text{th position}\} = \varphi p_i + \theta; \quad \varphi = 1 - \theta.$$

$$P\{E_j \text{ in the } (k+1)\text{st position given } E_i \text{ in the } k\text{th position}\} = \varphi p_j, \quad j \neq i,$$

for all $i,j=1,2,\cdots,s$. This means that we assume the persistence factor, θ, is the same for all pairs of like events. Now θ is some parameter and it is usually necessary to estimate it from the data. It is shown that T, the number of transitions, from one event to a different event, in the sequence is a sufficient estimator for θ if $p_i = 1/s$, $i=1,2,\cdots,s$, and it is accordingly suggested that an adequate estimator for θ when the probabilities, p_i, are unequal will be

$$\tilde{\theta} = 1 - \frac{T}{(n-1)(1-P_2)}$$

where

$$P_2 = \sum_{i=1}^{s} p_i^2.$$

We have also

$$\text{var } T = (n-1)\varphi\left[(1-P_2)(P_2\varphi+\theta) + 2(P_3-P_2^2)\right] - 2(P_3-P_2^2)(1-\theta^{n-1})$$

(from which var $\tilde{\theta}$ is immediate), with

$$P_3 = \sum_{i=1}^{s} p_i^3.$$

$\tilde{\theta}$ is asymptotically normally distributed. Given the probabilities $\{p_i\}$, therefore, it is possible to test a hypothesis about θ. Further, it is suggested that if there are two sequences, the first yielding an estimate $\tilde{\theta}_1$ (with associated probabilities $\{p_i\}$) and the second yielding an estimate $\tilde{\theta}_2$ (with associated probabilities $\{\mathcal{p}_i\}$), then a test criterion for the equivalence of the persistences in the sequences might be

$$\tilde{\theta}_d = \tilde{\theta}_1 - \tilde{\theta}_2$$

with

$$\text{var } \tilde{\theta}_d = \text{var } \tilde{\theta}_1 + \text{var } \tilde{\theta}_2$$

and $\tilde{\theta}_d$ normally distributed. It will be necessary to use an estimate of θ in var $\tilde{\theta}_d$ in order to carry out the test of significance and the estimate

$$\tilde{\theta} = \frac{\tilde{\theta}_1(n_1-1)(1-P_2) + \tilde{\theta}_2(n_2-1)(1-\mathcal{P}_2)}{(n_1-1)(1-P_2) + (n_2-1)(1-\mathcal{P}_2)}$$

where n_1 and n_2 are the numbers in the first and second sequences, respectively, and

$$P_2 = \sum_{i=1}^{n_1} p_i^2, \qquad \mathcal{P}_2 = \sum_{i=1}^{n_2} \mathcal{p}_i^2.$$

As an example of how to test a hypothesis about θ (or $\varphi = 1 - \theta$) wind data from Batavia for January, 1933 was given. Using past records the probabilities of the various wind directions in January at Batavia may be taken to be:

Direction.	S	SE	N	NE	N	NW	W	SW	C
Probability.	0.02	0.01	0.04	0.05	0.15	0.41	0.30	0.01	0.01

where C stands for calm. These are the $\{p_i\}$. The 31 actual wind directions for January, 1933 were

W, NE, SW, N, N, N, NW, W, N, NW, NW, N, NW, W, N, N, N,

N, NW, N, W, W, W, W, W, E, NW, W, W, NW, W.

The number of transitions $T = 19$ which gives an estimate

$$\tilde{\theta} = 0.114$$

with

$$\operatorname{var} \tilde{\theta} = \varphi(0.05964) - \varphi^2(0.03333) - 0.000433$$

if the term in θ^{30} is neglected. To test the hypothesis $\theta = \theta_0$ (or $\varphi = \varphi_0 = 1 - \theta_0$) we calculate the criterion

$$\frac{\tilde{\theta} - \theta_0}{\sqrt{\operatorname{var} \tilde{\theta}\big|_{\theta=\theta_0}}}$$

and refer to normal tables. Thus if $\theta_0 = 0.1$ we have

$$\operatorname{var} \tilde{\theta} = 0.0262, \qquad \sigma(\tilde{\theta}) = 0.162$$

and the test criterion is

$$\frac{0.014}{0.162} = 0.09$$

which is clearly not significant.

A preliminary numerical investigation into wind records with an estimate of $\tilde{\theta}$ calculated for each month of the year showed that the values of $\tilde{\theta}$ were approximately cyclical, which might be a reasonable thing to expect. It is not possible however to say whether or not this effect is a real one without considerable more numerical analysis with possible modification of the mathematical model proposed.

13. Further Research Work Possible on Persistehce. The writer was informed that the work on persistence fell only marginally within the scope of the contract and consequently further work on the topic was abandoned. Since the termination of the contract I have taken up the research again and have obtained some results which will possibly eventually appear in a scientific journal. Briefly the further work is as follows: To fix ideas let us think of persistence in wind direction with the compass as a clock face. In the model proposed in AFCRL-62-496 we wrote

$$p_{ji} = P[E_j | E_i] = \varphi p_j, \qquad i=1,2,\cdots,s \text{ but } i \neq j.$$

This means that the weights used for the probabilities are the same and equal to φ, and do not depend on the fact that the states E_i and E_j may be far apart. Thus for example if the state E_i is equivalent to a North direction of the wind, then

the probability of a wind direction NE given a previous North direction will be φ times the probability of a NE direction, and the probability of a S wind direction given a previous N direction will be φ times the probability of a S direction. It would appear more realistic to introduce a system of weights for the transitional probabilities dependent not only on φ but on the angular displacement of the wind direction assigning a bigger weight (say) to the transition from N to NE than to the transition from N to S, and this I have done. Whether such a model will be appropriate for analyzing the data of experience can only be tested from actual observational material.

14. General Remarks. The idea of building a stochastic model to describe the observations made on a particular phenomenon is obvious enough although I have not seen anywhere the approach to model building used here and came to it by myself. It has however probably been used in different guises many times before. Granted that the actual dynamics of the various aspects of the weather--rain, pressure, wind, etc.--are difficult, the approach, backward as it were, from what is actually observed may be helpful in offering a guide to the dynamic investigations. As the writer sees it one should build a tentative model, checking the validity of the model at each stage from observational data. Once the plausibility of the model as descriptive of the observed phenomenon is established, the mathematical consequences of the model can be worked out. There appears to be a large field of research in investigating the general models which could be proposed.